Working Paper Number 31

Anti-poverty Policy: Screening for Eligibility Using Village-level Evidence

Ruhi Saith[1] and Barbara Harriss- White

In the context of targeting of state transfers based on income poverty lines, this study is concerned with the identification of households that may have been wrongly included in the target group. To this end, we investigate the relationship between self-declared private income and some 478 household variables obtained in a village level survey. We use class probability tree analysis which is a non-parametric multivariate method. Relationships are expressed as easily interpretable rules that give combinations of the important features that characterise the "poor" households (income declared below the income poverty line) and the "non-poor" (income declared above the income poverty line), rather than as mathematical equations as in previous regression based analyses. Approximately 20% of the households that declared income so as to be classified "poor" were found to have feature combinations which were similar to those characterising "non-poor" households. These cases would thus be worthy of further investigation for distortion of income, before being considered eligible for any transfers.

January 2000

International Development Centre, Queen Elizabeth House, 21 St Giles, University of Oxford, OXFORD OX1 3LA

[1] To whom correspondence may be addressed. E:mail: ruhi.saith@wolfson.ox.ac.uk

Acknowledgements

This project was funded by a grant from the Department for International Development, UK Government The authors however are responsible for the results and conclusions of this paper. The authors thank Mr N. Narayanan, I.A.S. for his suggestions and support to the project in its embryonic form; Dr S Janakarajan and his team at the Madras Institute of Development Studies (MIDS) who collected the Census data; Prof. S.Subramanian (MIDS) for an updated poverty line; U. Archana (MIDS) for answering queries related to the data; Dr. Ashwin Srinivasan at the Oxford University Computing Laboratory for advice on class probability tree analysis method and very profitable discussions; Diego Colatei for most useful comments on the initial draft and computing assistance; Prof. Frances Stewart for critical comments and Prof. Quinlan for answering queries related to See5. R.S. was supported during part of the work by a Wingate Scholarship.

Approximate length of article, including footnotes and references, but excluding Appendices = 9000 words

Anti-poverty policy: screening for eligibility using village-level evidence

1. Introduction

In the late eighties, the Indian economy was beset with unsustainable deficits in the balance of payments and the domestic fiscal balance, high and rising debt-service ratios and inflation. In 1991 India took loans from the International Monetary Fund and the World Bank which carried the usual conditions of stabilisation and structural adjustment including public expenditure cuts. Given the strict control on public borrowing and pressure to reduce public expenditure, welfare transfers or targeted subsidies are the main policy instruments being used to ensure that the poorest who are known to depend most on the state are not the worst-hit[2].

In India, as elsewhere, those considered eligible for anti-poverty benefits are usually identified by using information related to income. This is despite the general acceptance of poverty as a multi-dimensional concept involving aspects as varied as income or consumption, power and social exclusion, functionings (e.g. health and education), vulnerability and livelihood unsustainability (Maxwell, 1999). While income will be used in our study as the variable to 'target' households, it should not be considered an implicit acceptance of income as an appropriate dimension by which to identify the poor. Rather, we use the available income-based data to introduce a new method of analysis of screening criteria which may find application in a range of other policy-oriented poverty studies, as well.

In India, the standard dimension of poverty used for targeting is private income, with people below a pre-defined poverty line being eligible for a variety of social transfers[3].

---

2 Lipton and Maxwell, 1992 and Lipton and Ravallion, 1995, trace the paradigms in international development in recent decades. The dominant paradigm in the 70's comprised industrialisation, infrastructure and re-distribution with growth. Poverty reduction strategies were directed towards basic needs and integrated rural development. In the 80's however, the focus shifted to structural adjustment with an emphasis on state compression and an expansion of markets. Towards the late 80's, this was modified to Adjustment with a Human Face (Cornia et al, 1987), protecting the social sector. The main principles of the resulting 'New Poverty Agenda' as read from the publications of the World Bank, the United Nations Development Programme and the United Nations in the early 90' have been succinctly summarised by Lipton and Maxwell, 1992. Of importance here is the emphasis on ensuring that the poorest are not the worst-hit, by targeting subsides or welfare transfers.

3 Two methods commonly used in developing countries to compute poverty lines are (a) the food energy intake method and (b) the cost of basic needs method (Wodon, 1997a). In the former, the level of consumption or income at which households would be expected to satisfy the normative nutritional requirement is calculated. Poverty lines are set taking this computation into account (Dandekar and

In addition to other problems involved in the measurement of income, discussed widely in the literature, income may often be deliberately and opportunistically understated to ensure inclusion in a set known to be eligible for state transfers. Attempts have therefore been made to improve income based targeting by using proxy indicators which correlate with income but are reliable, non-fudgable and practically easy to obtain.

The relationship between income and other variables which has been extensively explored in the literature related to 'poverty profiles', has provided some pointers towards proxy variables. Most quantitative studies have been univariate (looking at the relationship between income or other wealth indicators and single variables, e.g. family size, the dependency ratio, the education of household members or gender of head of household)[4]. More recently a few 'poverty status' analyses use a multivariate probit or logit framework in which a dichotomous poverty variable is regressed on independent variables such as region, household size and composition and assets ownership[5]. While such multivariate techniques can examine a large number of features simultaneously, the mathematical expression of the relationship between household variables and income makes implementation of policy using such complex results for identifying target households difficult. Further-regression based studies are limited in their ability to capture the non-linear relationships which could exist between income and other variables (e.g. income earned and age). In this paper, we use the 'class probability trees' analysis technique, which overcomes these drawbacks. This works by analysing data related to a sufficient number of cases belonging to different classes or groups (e.g. 'above' and 'below' a poverty line i.e. 'non-poor' and 'poor'). The pattern of variables characterising each class is discovered. Complex inter-feature

---

Rath, 1971; Greer and Thorbecke, 1986 and Paul, 1989). Under the cost of basic needs method, in addition to the cost of a food basket that enables the household to meet the normative nutritional requirement, an allowance for non-food consumption is taken into consideration. Poverty lines are set taking into account this combined computed cost (Ravallion and Bidani, 1994; Ravallion and Sen, 1996).

[4] See for instance, chapters by Sanyal and by Gaiha in Srinivasan and Bardhan (eds.), 1988; by Krishnaji in Rodgers (ed.), 1989; by Jayaraj and Subramanian in Harriss-White and Subramanian (eds.), 1999; and the review by Lipton and Ravallion in Behrman and Srinivasan (eds.), 1995.

[5] See Gaiha, 1988, Glewwe and Kanaan, 1989; Ruggeri, 1997; Wodon, 1997b and the studies referred to by Baulch and McCulloch, 1998.

interactions are automatically taken into account and results are expressed as rules that are easy to understand and apply.

An accurate understanding of the relationship between income and other variables has the potential to (a) identify variables that correlate with income and can be used as proxy variables (b) aid practical targeting by corroborating and refining existing criteria for the selection of income poor households and (c) yield insights into factors either influencing or associated with the income of the household and the causes of income poverty. This paper is focused on (b). An investigation aimed at (a) which has been commonly attempted in the literature requires the availability of data in which the income dimension is known to be accurate (Glewwe and Kanaan, 1989).[6] For identification of proxy variables, reliable estimates of income are essential. In the census data used here however, we cannot be sure that declarations of income by some households have not been understated.[7] We cannot therefore use class probability tree analysis to discover a pattern of features, characterising 'poor' and 'non-poor' groups with the view to identifying proxy variables as mentioned in (a) above. Instead, we concentrate on contributing to (b). In our data, amongst households that have declared their income as under the poverty line, we would expect certain households (i.e. the understaters) to be in reality, non-poor. The pattern of features characterising 'poor' using this data would therefore not be entirely reliable. On the other hand, households that have declared their income as above the poverty line, are unlikely to contain any 'poor' households[8]. The pattern of features characterising the 'non-poor' households, obtained using this data can thus be considered fairly representative of the

---

[6] Glewwe and Kanaan, 1989, use expenditure data instead of income data for the following reasons. First, as income can only generate welfare if used for consumption purposes, expenditures being closely tied with consumption levels are more appropriate. Second, expenditures are more accurate indicators than income for people whose income fluctuates from year to year (or as in case of farmers, season to season). Third and 'more compelling' is that income-data is often under-reported by respondents, due to fears of taxation, but expenditure data obtained by asking many specific questions is assumed to be less likely to be under-reported (Glewwe and Kanaan, 1989, p 11).

[7] When income related information is being collected for means-testing, households are likely to understate their income in order to ensure inclusion in a low-income group considered eligible for state transfers. When such information is collected for other purposes, e.g. census, some people may still understate income considering the possibility that the records may later be used for means-testing or for liability for tax.

[8] Sometimes, income may be deliberately overstated because people are ashamed of a low income It is quite unlikely however, that in census data, where the information is being collected by an outsider, households would deliberately overstate their income.

households that are indeed 'non-poor'. Given such information, we investigate the ability of the class probability tree analysis method to identify reliably households that have declared their income as being below the poverty line (i.e. "income poor") but are actually above. We expect the pattern of such "understaters" to resemble that of the 'non-poor' households. This as explained, earlier is a pattern that can be reliably obtained from our data.

The results reported here are not discussed with a view to contributing to (c), the causes of poverty, which would require further analysis and additional data.

2. Class probability trees and rules

Tree techniques have earlier been explored successfully in diverse analyses such as guiding the landing of the Space Shuttle (Michie, 1984), evaluating the credit-worthiness of clients applying for a credit card (Michie, 1988), location of primary tumours (Clark and Niblett, 1989) and guiding the selection of embryos in *in vitro* fertilisation treatment (Saith *et al*, 1998).

2.1 Introduction to class probability trees and rules[9]

Like traditional methods of multivariate statistics, the class probability tree method can analyse a large number of features simultaneously. Unlike traditional methods, which only capture linear relationships between features, complex inter-feature interactions are automatically taken into account. Very importantly, results are expressed as rules and are easier to understand and apply than mathematical equations.

The class probability tree analysis technique works by analysing data related to a sufficient number of cases. The cases belong to different classes or groups and the pattern of variables characterising each class is discovered[10]. See Table I which shows data related to cases belonging to two groups and described by 4 variables.

---

[9] This introduction has been largely taken from a paper describing an application of class probability trees and rules to the analysis of In Vitro Fertilisation data (described in Saith *et al*, 1998).

[10] Note that the word 'class' throughout this paper, is used in the statistical sense of the word, rather than to mean 'class' as used in Social Sciences.

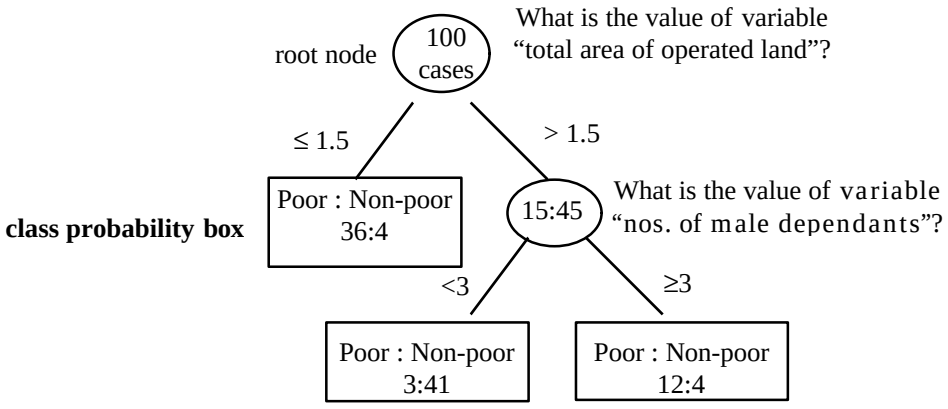
Table I Example of cases for constructing a class-probability tree

Case (household) nos.	Total area operated land (acres)	Nos. of male dependants	Nos. of females working	Cash balance (Rs)	Class
1	5	2	0	20,000	Non-poor
2	3	1	0	17,000	Non-poor
3	1	2	1	1000	Poor
4	2	4	3	200	Poor
5	10	3	1	40,000	Non-poor
…	…	…	…	…	…
100	18	2	1	10,000	Non-Poor

The pattern-class relationships are initially expressed as trees which are then re-expressed as a set of easily understandable statements or rules.

As can be seen in Figure 1 the tree starts as a 'root node', with a set of cases that are to be used to construct outcome-predicting rules. In the example here, these are 100 households referred to together as the 'training set'[11]. These cases are known to belong to mutually exclusive classes (here classes 'Poor' and 'Non-poor'). Each training case of a known class is described by its variable values (here, 4 variables related to each household). The training data is analysed by the class probability tree program and patterns of variables characterising and discriminating the 'Poor' and 'Non-poor' classes are identified. The computer program used here to conduct this task is See5 (Quinlan, 1998).

Figure I Hypothetical class probability tree



Details of tree-extraction methods can be obtained in Quinlan, 1993. Briefly, See5 looks at the root node to determine if all cases at this node belong to the same class. If they do not, as in the example in Figure I, it then grows branches from this node using the variable that best sorts the households (here, variable "total area of operated land" in Figure I) into distinct groups. The test that determines which variable is the 'best' considers the ability of the variable to sort into groups that contain a high proportion of households belonging to the same class. The ideal group has either an all 'Poor' or an all 'Non-poor' outcome. Here different values for the variable "total area of operated land" (1.5 and > 1.5 acres) gave two groups. One clearly had a high proportion of batches belonging to the class 'Poor' (36 out of 40) and the other to the class 'Non-poor' (45 out of 60). Following the split at the root node, the resulting nodes are examined. Further splits may take place (e.g. "nos. of male dependants" in Figure I) or the tree may terminate at a node. Termination happens in two situations. First, if all cases at a node belong to the same class and second if the test indicates that the gain by further splitting is unlikely to add to the overall discriminatory power (as here with the variable total area of operated land ≤ 1.5). Such a terminal node is declared a 'leaf' or 'class probability' box. In Figure I the proportion of cases belonging to each class in that box are shown. When used for outcome-prediction, this box provides an estimated probability of whether a new household would be predicted as belonging to the class 'Poor' or 'Non-poor' (hence the name class probability trees). This is achieved for a new household by checking the condition at the top of the tree and working down the branches depending on which branch its features satisfy until a 'class probability' box is reached.

The variable pattern-class relationship expressed in the tree in Figure I can also be written as a set of rules as shown below:

Rule Set

Rule 1

If

Variable "total area of operated land (acres)" has a value less than or equal to 1.5

Then

---

[11] These cases are referred to as 'training' cases as the analysis program 'trains' on the information provided by them and extracts patterns which can then be used to classify new cases.

Probability of 'Poor' is 90 % (36/40) and probability of 'Non-poor' is 10% (4/40)

Class prediction taken as **Poor**

Rule 2

If
Variable "total area of operated land" has a value greater than 1.5 and
Variable "nos. of male dependants" has the value less than 3
Then
Probability of 'Poor' is 7% (3/44) and probability of 'Non-poor' is 93% (41/44)

Class prediction taken as **Non-poor**

Rule 3

If
Variable "total area of operated land" has a value greater than 1.5 and
Variable "nos. of male dependants" has a value greater than or equal to 3
Then
Probability of 'Poor' is 75% (12/16) and probability of 'Non-poor' is 25% 4/16)

Class prediction taken as **Poor**

Each rule contains class probability estimates expressed as percentages. These have been obtained from the proportion of cases in each class in the 'class probability' box in the tree. In this example, the class label ('Poor' or 'Non-poor') attached to the rule is that of the class that has the higher probability. Note that each rule bases 'class prediction' on a majority-vote principle. That is to say, where less than 50% 'Poor' is predicted, the class prediction is 'Non-poor', and otherwise 'Poor'. The 50% prediction criterion is convenient, particularly given the structure of the rule-building See5 package, but it is to a degree arbitrary.

3. Data

3.1 Source of data

Data used for this study were obtained from a village-level census conducted in 1994 by a team from the Madras Institute of Development Studies, India. The census was conducted to update base-line data collected previously for the same villages in 1973-74 and 1982-84. Of the 11 villages for which data were collected in the most recent census in 1993-94, our analysis is restricted to data from the populations of 2057 households from the following eight villages: Vegamangalam, Sirungathur, Duli,

Vengodu, Vayalur, Meppathurai, Amudhur and Kalpattu located in the north of Tamil Nadu[12].

The variables in the census data were not deliberately selected to focus on poverty, but at the same time were not arbitrary. Census variables were selected in 1972 to give the best possible economic and social characterisation of agrarian households (Farmer *et al*, 1997). The list of variables included in the census was modified 10 years later to cope with the growing importance of the non-farm economy (Hazell and Ramasamy, 1991). Other additions to the recent census in the 90's included data related to social welfare, access to a range of state interventions, gender and water management. Data contained in the census data base are thus voluminous. The use of such a data-base to focus on poverty and to examine relationships between income and other variables characterising households as here, has the distinct advantage over previous studies of the incorporation of a large number of variables which have not been preselected.

3.2 The construction of variables

The census data for the eight villages contains variables collected under 10 different headings as follows:

Table II Groups of variables describing each household

No.	Group
1	Demographic and Occupational profile
2	Migration details
3	Housing condition
4	Welfare (education and reproductive health)
5	Land holding status
6	Cropping pattern 1992-93
7	Agricultural assets
8	Irrigation status
9	Non-agricultural assets
10	Liabilities

Approximately 400 variables related to each household were collected pertaining to the 10 groups. Not all of the variables however, were in a form that allowed direct

[12] A pilot study investigating the feasibility of application of the class probability tree method for poverty analysis had been conducted on data for the three other villages of Nesal, Vinayagapuram and

inclusion in the current study. This information had to be transformed so that each household would have the same number of variables, irrespective of the number of members. Thus instead of the age or education of each member, variables that specified the number of males or females in the household that belonged to a particular age group (see variables. 3 - 22 in Appendix I) or had a particular level of education (see variables. 27 - 44. in Appendix I) were obtained from the original information and included in the analysis. Similar transformations were made with respect to data on the occupation, receipt of social welfare schemes, illness/handicap, migration details and chit loans belonging to each member of the household. In some instances even if differences in particular components may not be important characterising factors of the household class, differences in the total might be significant To account for this possibility, some new variables were constructed by combining the values for a group of individual variables to give the totals. (for example, total area/value of land quality types owned, leased, mortgaged or rented - see variables in Appendix I). For the same reason, values were pooled for crops of different types grown on irrigated or un-irrigated land, the monetary value of assets of different types owned by the household, liabilities owed to different institutional or non-institutional sources and details of the chit loans. Thus, in addition to 273 variables included in their original form, an additional 205 new variables were constructed. Appendix I details all the 478 variables used in the analysis and also mentions whether the variable was included in its original form or was obtained by computation of the available census data.

4. Method of analysis

4.1 Variables included in the study

478 variables relating to 2057 cases were included in the analysis.

4.2 Program used to conduct the analysis

The tree program See5 was executed under the Microsoft Windows 95 Operating system. As with most tree-based analyses, it is capable of analysing data related to a

---

Veerasambanur (Saith, 1996).

large number of cases and including a large number of variables. Details can be found in Quinlan, 1998.

4.3 Obtaining Rules

Data describing 2057 households from eight villages were used for this analysis[13]. Each household was classified as belonging either to the class of 'Poor' or to that of 'Non-poor' using a poverty line defined by the Central Government of India.[14] A household was classified 'Poor' if per capita income per month was below the poverty line (defined at Rs. 195.31) and classified as 'Non-poor' if the income was greater than or equal to this amount. Roughly half (1033 households) were thus classified 'Poor' and the remaining half (1024), 'Non-poor'. Class probability tree analysis was conducted on the data to obtain rules giving the variables characterising 'Poor' and 'Non-poor' classes.

Our hypothesis is as follows: if households have declared their income correctly, we would expect a consistent relationship between income and the values of the other variables describing them. The pattern of correlates of households that have on the other hand, for one reason or another, wrongly declared their income as being below the poverty line when they are actually above, would be expected to be more closely related to that of households that have an income above.

---

[13] Data were available for 2067 households. Seven households for which the income variable was not available were however excluded. In addition, 3 households that showed an income of zero despite the possession of own land, raising the possibility that the income may have been entered as zero by mistake, were also excluded.

[14] For different methods that may be used to obtain poverty lines and the advantages, disadvantages and problems associated with these or the different poverty indices used to aggregate information, see Lipton. and Ravallion 1995.

In the Indian context, the poverty line is either that decided on by the Central Indian Government's Planning Commission or one announced by State governments. In this study the poverty line along the lines recommended by the Central Government of India was used. The poverty line for rural Tamil Nadu for 1973-74 in terms of monthly per caput consumption expenditure is Rs.45.09 per caput (Planning Commission, GOI, 1993). This had to be updated to 1992-93 prices as household income data collected in the census pertained to 1992-93. Subramanian updates this as follows. The Consumer Price Index of Agricultural Labourers (CPIAL - available in Indian Labour Journals) for Tamil Nadu for 1992 -93 with 1973-74 as base was calculated to be 424.38 (the 1992-93 CPIL for Tamil Nadu with 1960-61 as base is 1027, while for 1973-74 it is 242. Implicitly therefore the CPIAL Index for 1992-93 with 1973-74 as base is 1027/242 = 424.38). Tamilnadu's rural poverty line at 1992-93 prices was then estimated to be Rs.195.31 per month (= 45.09 * 4.2438).

4.3.1 Training and test sets

The data of 2057 households was split randomly into 2 groups. One group consisted of 1500 cases (i.e. the "training set") and the other (the remaining 557) comprised the "test set"[15]. The training set is the sample from which sets of outcome-predicting rules are constructed. The test set is used to test these rules to see how well they predict the Poor/Non-poor outcomes of a new set of cases. This is standard practice adopted to evaluate the usefulness of such techniques (Quinlan, 1998).

4.3.2 Error rate of rules obtained

The performance of a given set of rules derived from a class probability tree, when classifying N cases can be summarised by a 2×2 (two by two) table. This is shown in Table III, together with the meanings of the entries. The entries have been labelled taking into account the context of our study in which the data simulate the scenario of a targeted anti-poverty scheme which involves 'beneficiaries' and 'non-beneficiaries', of benefits, based on declarations of income.

[15] The split was done by using a computer program written by Dr Ashwin Srinivasan, such that the proportion of 'Poor' and 'Non-poor' households within the training and test set were similar to that in the full data set of 2057 households i.e. approximately 'Poor': 'Non-poor' = 50:50.

Table III 2 × 2 table

Predicted class

Declared class	Poor	Non-poor	
Poor	CB	IB	N1
Non-poor	IN	CN	N2
	N3	N4	N

CB = Correct Beneficiaries
IB = Potentially Incorrect Beneficiaries
IN = Potentially Incorrect Non-beneficiaries
CN = Correct Non-beneficiaries

The declared class is the class to which a case belongs i.e. the class assigned based on the income declared by the household. The predicted class is the class assigned to a case by the rule (i.e. pattern of variables) it satisfies. CB, or correct beneficiaries are the number of cases that declared themselves as having an income below the poverty line, i.e. being 'Poor' and are assigned (i.e. predicted or classified by the rules) as such. IN, or potentially incorrect non-beneficiaries are the number of cases that declared themselves as having an income above the poverty line i.e. being 'Non-poor' that are classified by the rules as 'Poor'. Thus the pattern of variables characterising these households resembles that of 'Poor' although they have declared themselves as belonging to class 'Non-poor'. CN, or correct non-beneficiaries are the number of 'Non-poor' cases correctly classified as such. IB, potentially incorrect beneficiaries are the number of self-declared 'Poor' cases classified here as 'Non-poor'. The pattern of variables characterising these households thus resembles that of 'Non-poor' although they have declared themselves as belonging to class 'Poor'. These cases are the focus of this paper.

N1 (CB+IB) is the total number of cases that belong to the class 'Poor' i.e. those households that have declared their income as being below the poverty line. N2 (IN+CN) is the number of cases that belong to the class 'Non-poor' i.e. those that

have declared their income as being above the poverty line. N3 (CB+IN) is the total number of cases classified by the rule set as 'Poor' and N4 (IB+CN) as 'Non-poor'. N gives the total number of cases in the training set. From Section 2.1, we known the manner in which the trees and rules are constructed and the class label attached to the rule. The error rate of the rules is then estimated by the fraction of cases incorrectly classified, that is, (IB+IN)/N. An assessment of how good the ability of classification of the rules obtained is can be made by comparing this error with the error made by a very simple classifier which would predict that every new case would belong to the most common class in the training data. Our training data of 1500 cases had approximately 744 'Poor' households and 756 'Non-poor' households. A majority classifier would thus classify any new case as 'Non-poor'. For the test set which has 557 cases with 289 'Poor' and 268 Non-poor', the majority classifier would classify all cases as 'Non-poor'. 289 of these are however 'Poor'. The error rate of the classifier would thus be 52%. The pattern of rules constructed by See5 would thus be judged by comparing their error rate against the majority classifier error rate of 52%.

These error rates are of two kinds: 'apparent' and 'predictive'. Apparent error is the error rate secured when the calculations above are performed on a 2x2 table obtained from classifying cases in the training set (that is, the cases used to construct the rules in the first place). Usually the apparent error of a set of rules will be more optimistic (that is, lower) than its predictive error i.e. its ability to predict the outcome of new cases. This is because the rules may be 'over-fitting' the training data[16]. An estimate of the predictive error can however, be obtained by using the rules to classify new data i.e. the test set. For this estimate to be reliable however, a very large number of cases are essential in the 'training' and 'test' sets. Since our data set is not very large, an alternative procedure to obtain an unbiased estimate of predictive accuracy is the procedure of cross-validation (Weiss and Kulikowski, 1991). In this procedure, the training set (N) is divided so that a few cases (k) are 'left out'. This sub-set serves the role of new data. The rules obtained by training on the remaining cases (N-k) are

[16] The tree (and resulting rule set) is obtained by analysing cases in the training data. Classification of training data using this Rule Set may give low errors because rules may have been specifically constructed to characterise particular cases. The pattern is highly specific to the training data and is said to 'overfit'. The pattern may thus not hold good on new data.

tested for their classification ability on this sub-set, also referred to as test set. This procedure is repeated by 'leaving out' a different group of k cases each time selected at random. Thus each case in the sample is used as a test case and each time most of the cases are used for training. The estimate of predictive error of the rule set obtained by training on all the cases is then the average of the predictive error for each sub-set sample calculated from a 2x2 table, similar to that shown above.[17] Such estimates of predictive error are free from bias of the kind found in apparent error rates and give the truer picture.

If adequate data are available, a further, and independent, unbiased estimate of predictive error may be obtained when the rule is used to classify a separate test set of new cases drawn from the same data source. This is not essential, and just helps to confirm the estimate obtained by cross-validation. Note that this test set consisting of new cases differs from the test set used to obtain cross-validation results, which were sub-sets of the training data itself.

4.3.3 Tree and rule construction

The data for the training set of 1500 households with 744 belonging to the 'Poor' class and 756 to the 'Non-poor' class were analysed using See5. The rules were obtained by analysing data related to 478 variables at default See5 settings.[18] Details of the tree construction process and that of converting the trees to rules (which indicate the variables characterising and discriminating 'Poor' households from 'Non-poor' households) are in Quinlan, 1993.

---

[17] Consider a training set with N cases (usually 2/3rd of data available). The test set (remaining1/3rd of the data) will have less than N cases. In cross-validation, a number of sub-sets play the role of the 'test' set. Each of these sub-sets is smaller than a single test set. As different parts of the training data are sequentially used as 'test' sets however, the effect is that of using a test set of size N.

[18] We have used the default settings which are expected to give a reasonably low error rate. Obtaining the lowest possible error rate would however require experimentation with a systematic variation of the parameter settings. Some researchers have investigated methods to do this (see e.g. Kohavi and John, 1995).

5. Results and discussion

The results are presented in the following sequence. The predictive accuracy of the Rule Set obtained at default settings is estimated by a 'leave 10 out' cross-validation. The 2×2 table representing this performance is given first (Table IV). This is followed by a 2×2 table (Table V) representing the performance of the Rule Set on the test set of 557 households.

Table IV 'leave 10 out' cross validation 2×2 table on training set

		Predicted class		
		Poor	Non-poor	
Declared class	Poor	586	158	744
	Non-poor	202	554	756
		788	712	1500

Estimated predictive error of the rule set = (158 + 202)/1500 = 24%
The proportion of households declaring income so as to belong to class 'Poor and classified 'Poor' i.e Correct beneficiaries (CB) = 586/744 = 79%
The proportion declaring income so as to belong to class 'Poor' and classified 'Non-poor' i.e. Potentially incorrect beneficiaries (IB) = 158/744 = 21%

The estimated predictive error at 24% is significantly lower than that of the majority classifier error rate (52%). These results also indicate that of the 744 households that declared their income as 'Poor, 586 (79%) are classified by the rules as 'Poor. This suggests that the pattern of variables characterising these households is consistent with this low income. The remaining 158 (21%) households however are classified by the rules as 'Non-poor'. The pattern of variables characterising these households is more similar to households which declared their income above the poverty line. These results are confirmed by using the previous rules to classify the remaining test data in which 78% households were classified in the former group and 22% households in the latter (see Table V).

Table V 2×2 table obtained on the test set

Declared class \ Predicted class	Poor	Non-poor	
Poor	225	64	289
Non-poor	71	197	268
	296	261	557

Estimated predictive error = (64 + 71)/ 557 = 24%
The proportion of households declaring income so as to belong to class 'Poor' and classified 'Poor' = CB = 225/289 = 78%
The proportion of households declaring income so as to belong to class 'Poor' and classified 'Non-poor' = IB = 64/289 = 22%

The Rule Set (detailed in Appendix II) characterising 'Poor' and 'Non-poor' households suggests that of the 478 variables analysed, about 40 variables have a close relationship with the income of the household. As an illustration, two of the most reliable 26 rules are presented here.[19].

[19] The full Rule Set comprising 26 Rules is presented in Appendix II. Of these 10 Rules characterise "Poor" households and 16 Rules "Non-poor" households. Only one rule in each group has been presented in the Results section as an illustration. The factors taken into consideration when selecting one representative rule for each class have been a combination of the following:

A) **cover** - This indicates the number of cases in the training set that are characterised by the pattern of features presented in the rule. The higher the number of cases covered, the higher the likelihood of finding cases with similar features in new data. Besides the higher the number of cases covered by the rule, the more statistically reliable are the probabilities of prediction (i.e. B below) associated with each rule likely to be.

B) **accuracy**: This is the probability of each case covered by the rule, belonging to a class same as that indicated by the rule. The higher the value, the more accurately would the rule be expected to predict new data.

It would thus be reasonable to expect the value of A*B to be a good judge of the overall performance of the rule - the higher the value, the better. When the A*B value for two rules is similar or very close, the ease of interpretability of the rules may be used as an additional factor in judging the rule performance. The 'Non-Poor' rule presented above had the highest A*B value of the 10 Non-Poor rules. Amongst the 'Poor' rules, two rules with the highest A*B values had very close values. The rule with the slightly higher value was more complex (with 12 conditions) than the rule presented here (which was simpler with just 4 conditions).

Poor Rule

If
at least one (or more) female members is currently an agricultural labourer,
the value of total land leased out is equal to or under Rs.25000,
the gross overall production of all crops is equal to or under 1.8 metric tons and
the total amount of chit fund is equal to or under Rs. 15000
Then
Probability of 'Poor' is 77% and 'Non-poor' is 23%

Class prediction is taken as **Poor**

'Non-poor' Rule

If
at most two male members are dependant,
no male member is currently an agricultural labourer and
the gross overall production of all crops is greater than 1.8 metric tons
Then
Probability of Non-poor is 94% and Poor is 6%

Class prediction is taken as **Non-poor**

These 2 rules demonstrate the manner in which the results of class probability tree analysis are expressed. As expected, the features identified indicate that in rural households poverty is still largely defined by reference to the agricultural economy. The rule predicting 'Poor' households shows that features considered important in identifying 'Poor' households include number of female *agricultural labourers*, *value of total land leased out*, *output of crops* and *amount of chit fund the household subscribes to.*. This is supported by a cluster analysis of census data for the same period for three other villages in the region, showing the 'elite' cluster in all three villages to own as well as operate an average of at least 6 times more land than the 'peasant' cluster (Colatei and Harriss-White, forthcoming). The condition related to *leasing out land* (the value of total land leased out being equal to or under Rs.25000) appears counterintuitive as there is little tenancy at all in this region. Besides, it would be expected that most poor households would not own any land, let alone being capable of leasing out land. However 'Poor' households leasing out small amounts (up to a value of at most Rs. 25000) may be explained by 'reverse tenancy'. This is a contract where elderly females or households with sick members are forced to lease out their land to others, usually more able bodied. Janakarajan, 1996, discusses the impact of the changing irrigation scenario (of the disuse of tanks and a very high reliance on groundwater for irrigation), on land lease in the region. Lessors are found

to be poor farm households who do not own wells or whose wells have dried and are forced by circumstances to lease out their own land to adjacent better off well owners. It is further supported by the cluster analysis for 18 socio-economic variables from the census data of three other villages in the region. The analysis shows no land as being leased out by the 'elite' cluster for two of the villages, but average amounts of just 0.03 and 0.06 acres as leased out by the 'peasant' clusters. In the third village, however the average amount of land leased out by the 'elite' cluster (0.22 acre) is about twice that leased out by the 'peasant' cluster (0.13 acre). The *chit fund* mentioned in the last condition (the total amount of chit fund being equal to or under Rs. 15000) also needs some explanation. This is a Rotating Savings and Credit Association (ROSCA) common in the urban and rural informal finance sectors[20]. Although participation in chit funds is widespread, they are particularly common amongst the poor due to low transaction costs which increase their accessibility (for a discussion see Calomiris and Rajaraman, 1998 and Ardener, 1995). At first glance, the limit of the chit fund identified in the rule as Rs 15000 looks high for 'Poor' households (being above the annual income at the poverty line of an average household of size 4.5 members). But the condition also includes households which are not members of a chit fund. Of the 2064 households in the data, only 259 households (13%) participated in a chit fund. Of these, more than half had a total chit of less than Rs.15,000. About 60% of households in this sub-group had a total chit fund amount less than Rs 5000. About two thirds of this group with small sized chits, had the payment spread out over 20 or more instalments either seasonally (amounting to Rs.65 or less per month) or via monthly payments (amounting to Rs 250 or less per month). When viewed in terms of instalments therefore, it is entirely plausible for 'Poor' households to be characterised either by no participation in ROSCAs or participation in ROSCAs with small funds spread out over many instalments.

---

[20] Calomiris and Rajaraman, 1998 (p208) define a ROSCA = as follows: "…is a voluntary grouping of individuals who agree to contribute financially at each set of uniformly-spaced dates towards the creation of a fund, which will then be allotted in accordance with some prearranged principle to each member of the group in turn. Allotment is either through lottery (random ROSAs) or auction (bidding ROSCAs)." The chit fund referred to in this study is a bidding ROSCA. The total amount of money to be put in, the amount in each instalment, the number of instalments and interval of payment are all pre-decided. At each payment round, people who need money put in a bid for it. The bids are slightly lower than the total amount. The person with the lowest bid gets the money but continues to pay all the instalments.

The Rule predicting cases as 'Non-poor' indicates the commonest combination of features found characterising 'Non-poor' households. If a household has able-bodied and active workers with two or *fewer male dependant members* (i.e. either aged less than 15 or greater than 64 years), with none of the males in the household being employed as *agricultural labourers* , and is landed, with a *gross crop output* of at least 1.8 tonnes (or more), there is a high probability that the household has an income above the poverty line. These findings are supported by the cluster analysis mentioned above, where the dependency ratio was found to be consistently higher for the 'peasant' group compared to the 'elite' in all three villages[21]. Further, "Non-poor" households usually have a source of earned income (such as from wage work in the non-farm economy) other than from agricultural labour (Jayaraj, 1992 and 1996). Normally, they also have higher gross agricultural production than "Poor" households (Harriss-White and Janakarajan, 1997)[22].

We will not go further into the relationships identified between income and other variables in the entire Rule Set (which is presented in Appendix II), preferring to focus on the general pointers that can be obtained from results presented in this manner, irrespective of the specific variables used in this particular analysis.

In the results as presented in the 2×2 tables, the top row is of key importance. This corresponds to the households that have declared their income such as to belong to class 'Poor'. If the poverty line is used as the criterion for targeting, this is the group of major interest for targeting purposes[23]. The households that belong to the box (CB)

---

[21] The dependant age group in the cluster analysis was considered to be that aged less than 15 or greater than 60 years of age. This was necessitated because of the nature of data available for these villages. This result confirms the general conclusion of Lipton and Ravallion, 1995. Other studies, however find reduced dependency among the poor rather than the non-poor (see Ramu, 1988; the review in Harriss, 1992 and Rodgers conclusion that the relations between demographic or economic dependence and poverty in rural South Asia are 'weak', p15, 1989).

[22] The support provided by the findings of the cluster analysis showing the relationship between average amount of land operated by households in the 'peasant' and the 'elite' groups has already been mentioned earlier.

[23] We do not consider the lower row, i.e. households that have declared their income as being above the poverty line, as these households would not be included in the target group. We do not concern ourselves here with households that are in reality income poor but are not considered eligible for benefits (i.e. F errors mentioned in Cornia and Stewart, 1995) based on their declared income. We assume here that it would be unlikely that households that are income 'poor' would deliberately state their income so as to belong to class 'non-poor' and thus wrongly fail to receive benefits.

i.e. cases that declare their income so as to be classified as 'poor' and are predicted by the Rules too as 'poor' could be considered as "eligible". Hence these cases are referred to as 'Correct Beneficiaries'. All cases that fall in the box (IB), i.e. cases that declare income so as to belong to class 'Poor' but are predicted as 'Non-poor' would be worthy of further investigation before being considered eligible for any transfers Hence in the paper all cases belonging to the group IB, are henceforth referred to as '*Potentially* Incorrect Beneficiaries'. As the results presented in Tables IV and V indicate, about 20% of households are in this latter category[24]. The possible reasons for almost one fifth of households that had declared their income such as to belong to class 'Poor' having been found to have features similar to those characterising 'Non-poor' households are as follows:

- **Deliberate poverty distorters** Some households might falsely declare their income as below the poverty line when it is actually above in order to be considered eligible for perceived benefits
- **Poverty distorters due to underestimation** Some households are characterised by features suggesting a higher income than that declared by them e.g. high crop productivity, few dependants and no males employed in agricultural labour. The possibility exists that they may have underestimated their income by mistake, rather than with the intention to deceive. It is also possible that some households are characterised by features similar to those of other non-poor households although genuinely having a low net monetary income as declared. This could happen when any costly event such as medical treatment or ritual expenses is netted out of statements of income by a respondent.
- **Noise** These are mistakes made in the entry of data e.g. although the income may have been declared by the household such that it belongs to the class 'Non-poor', the house was wrongly classified as 'Poor'.
- **Prediction errors** The census data used in this study may not include all the appropriate variables. If some of the missing variables had been included, in the

[24], In an evaluation of six studies of the impact of the Integrated Rural Development Programme (IRDP), the average rich who were ineligible, but nevertheless got access to IRDP loans was 21% of households (Copestake, 1992).

census, it is possible, that other patterns of variables characterising households with an income below the poverty line, would have been identified, resulting in their classification by the rules as 'Poor'. It is also possible that See5 is not capable of capturing all possible patterns that characterise the households that declared their income as 'Poor'. Some outliers might thus have been left out and are mistakenly predicted as 'Non-poor'.

Using the class probability tree analysis technique, it is not possible to distinguish between cases belonging to the four categories above. What can be concluded is that all cases that fall in the box (IB), i.e. cases that declare income so as to belong to class 'Poor' but are predicted as 'Non-poor' would be worthy of further investigation so as to try and identify those that are strictly not eligible for any transfers[25] being poverty distorters (either deliberate or due to underestimation).

The methodology outlined above may have policy applications. Before the transfer of benefits to people considered eligible by virtue of a declared income below a set poverty line, the confirmation of eligibility might be reasonably desired. Households that have declared income so as to be classified "Poor" and are also predicted by the above methodology as being "Poor" would be directly considered eligible for transfer. On the other hand, all households identified as '*potentially* incorrect beneficiaries' would need further investigation for eligibility. The extent to which the technique of analysis presented above can be applied usefully will depend ultimately on the use of features that are easy to collect and less easy to distort than income.

We therefore repeated the analysis restricting the variables included to 75 'non-fudgable' variables identified from amongst the 478 variables (see Appendix I). These 'non-fudgable' variables cover physical assets, caste status, demographic data and data

---

[25] Recall that IB errors refer to households that have declared income so as to belong to class 'Poor' but are predicted as belonging to class 'Non-poor'. IN errors refer to households that have declared income so as to belong to class 'Non-poor' but are predicted as belonging to class 'Poor'. As with reasons for errors in IB, the reasons for errors in IN (approximately 28%) are over-estimation out of ignorance, noise, prediction errors and shame. Shame is the reverse of deliberate poverty distortion. Some households that have an income below the poverty line may feel ashamed and falsely declare a higher income. The pattern of variables characterising these households would thus be expected to resemble that of households belonging to class 'Poor'. This can certainly happen when income has been obtained as part of a larger general questionnaire as with the census data used in this study. If income has been obtained specifically for the purpose of targeting benefits or means testing however, it is quite unlikely that households would knowingly err on the side of declaring a higher income.

on health, variables which are most straightforwardly, visibly verifiable by an outsider. The variables that were excluded are largely those that describe aspects of the household economy where there is an inherent incentive to under estimate and/or where verification and cross checking is impossible or very costly. These latter are easily fudged. The other group that was excluded is the occupational vector which occupies an intermediate position in terms of 'fudgability'.

5.1 Results obtained with 'non-fudgable' variables

When the analysis is restricted to the 'non-fudgable' variables, the Rule Set characterising 'Poor' and 'Non-poor' households suggests that of the 75 variables analysed, about 36 variables have a close relationship with the income of the household. As an illustration, two of the rules from the rule set obtained by the analysis of 'non-fudgable' variables for the training data are presented here[26]:

Non - Poor Rule

If
the head of the household is male,
the number of female members in the working age group (aged between 15 - 64 years) is at most 1 and
the dependency ratio (dependants/working population) is at most 0.6
Then
Probability of Non-poor is 70% and Poor is 30%

Class prediction is taken as **Non-poor**

Poor rule

If
the household belongs to a scheduled caste,
dependency ratio(dependants/working population) is greater than 0.6,
the estimated monetary value of electrical pump sets owned is at most Rs 6668,
the estimated monetary value of oil engines owned is at most Rs 3333,
the value of agricultural implements owned is at most Rs 350,
the estimated monetary value of buildings owned in the village is at most Rs 35000 and
the estimated monetary value of business assets owned in the village is at most Rs 1000

[26] The criteria for selection of these rules are similar to those described earlier - i.e. rules with the highest Accuracy X Cover value.

Then

Probability of 'Poor' is 88% and 'Non-poor' is 12%

Class prediction is taken as **Poor**

We do not discuss the rules in detail but indicate below, the main features identified in the full rule set as characterising and discriminating the 'Poor' and 'Non-poor' households. Recall that these rules have been obtained by restricting the analysis to 75 non-fudgable variables. The rules presented earlier were obtained by analysis of the entire range of 478 variables.

1. **Caste status:** In the villages included here, households belonging to scheduled castes usually live in separate hamlets and are thus easily identified. This feature bears an important relationship to the income poverty status. The finding is supported by the cluster analysis of census data for the same period for the three other villages in the region which shows that most (more than 80%) Scheduled Caste households belonged to the 'landless peasant cluster' (Colatei and Harriss-White, forthcoming). Further, there is also much evidence from other research in India (Dreze and Sharma, 1998 in Palanpur; and elsewhere as reviewed in Agnihotri, 1997) confirming the high probability of income poverty among scheduled castes and tribes.

2. **Gender of the head of the household:** The gender of the household head was found in our study to be an important discriminatory feature of income 'Poor' and 'Non-poor' households. Further, Colatei and Harriss-White show that although exceptions existed, female headed households were most likely to be in the lower half of the asset distribution.[27]. Households composed solely of female members were always found to belong to the poorest cluster. The relationship between the gender of the household head and poverty is also well attested in the literature (see Dreze, Lanjouw and Sharma, 1998, for evidence of downward economic mobility amongst widows living without an adult male)

[27] In their cluster analysis, Colatei and Harriss-White take the gender of the respondent to indicate the gender of the head of the household. In the analysis conducted by us, the head was considered female in the following circumstances: (a) households with only female members; (b) households with male and female members, but with only females in the working age group (15-64). The head of the household was assumed to be male in all the other households.

3. **Demographic variables:** These include the type of family (nuclear or joint), and the gender and age distribution of household members as bearing an important relationship with the income poverty status of the household

4. **Illness:** The number of male and/or female members that are ill, handicapped or otherwise incapacitated from work matter in determining the income 'poor' or 'non-poor' status, a result confirmed both in National Sample Surveys (Subramanian and Harriss-White, 1999) and in the study of disability (Erb and Harriss-White, forthcoming).

5. **Assets:** 'Non-fudgable' assets identified as being useful in discriminating between the 'poor' and those that may in reality belong to the 'non-poor' group are the possession of electrical pump sets, agricultural implements and oil engines. Buildings and business assets possessed, inside the village, and thus easily verifiable by observation or by questioning neighbours, (as opposed to assets located outside the village) were also identified as important discriminatory factors.

6. **Variables describing the dwelling:** These include the material used for the construction of the floor, roof and walls of the dwelling, presence of an electricity connection, the kind of fuel used for cooking, the number of families occupying a single house and the village the family lives in. Similar use of dwelling to identify those in most need of assistance has more recently been explored by Bliven *et al*, 1997 who find that the poorest families in North Arcot in Tamil Nadu, were most likely to be residing in temporary shacks. Glaring differences in the kind of housing and electricity provision between the caste village and the Harijan colony in Iruvelpattu village in the South Arcot district in Tamil Nadu were also reported by Guhan and Mencher, 1983.

The results obtained for the non-fudgable variables by the 'class probability tree analysis' indicate the combinations of the different features mentioned above that characterise 'poor' households and 'non-poor' households. With 'leave 10 out' cross validation about one fourth of the 'poor' households are identified as being '*potentially* incorrect beneficiaries'[28]. These results are confirmed by using the rules to classify the

[28] Two by two table results not presented here for brevity.

remaining test data in which 74% households are also classified as 'correct beneficiaries' and 26% as '*potentially* incorrect beneficiaries'. The features identified in the set of 'Non-poor' rules could thus be potentially useful in a further assessment of the latter group of households for eligibility.

5.2 Results obtained with varying "costs"

Depending on political, social and financial considerations, the state may prefer to give benefits to a larger (or smaller) number of households directly (CB's) and investigate fewer (or larger) number of IB's. Such policy preferences can be taken into account by altering the cost of errors in the analysis. Rules presented above and in Appendix II were obtained by considering the cost of errors of both types as equal i.e. error of predicting households that declared income such as to belong to class 'Poor' as 'Non-poor' (IB or Type I error for a null hypothesis that a household is "Poor") was given an equal weighting to the error where a household that declared its income such as to belong to class 'Non-poor' was predicted as 'Poor' (IN or Type II error). As with any of the other multivariate analysis techniques, if making errors of one of the two kinds is considered worse, this could be taken into account while analysing the data, by incorporating a error cost. In decision theory, the cost of making such an error (an error thought to be worse) would be considered higher than that of the other. Consider as an illustration, the situation where a Type I error (classifying 'Poor' as 'Non-poor') is considered worse than a Type II (classifying 'Non-poor' as 'Poor') error. It is possible, using differentiated costs, to obtain less stringent rules (more general rules) to characterise 'Poor' households. A larger number of households will be classified as 'Poor', thus reducing the chances of a 'Poor' household wrongly being classified as 'Non-poor'. However, this also means that a larger number of 'Non-poor' households will also be classified as 'Poor'. Thus the number of Type II errors will increase although the Type I errors decrease.

Depending on which error is considered more or less costly, the proportion of households classified in the top row i.e. CB:IB or 'Correct Beneficiaries': 'Potentially Incorrect Beneficiaries' changes as shown in Tables VI and VII.

Table VI Effect of increasing cost of error IN greater than IB

Cost of error IB	Cost of error IN	Proportion CB:IB
1	**1**	**79:21**
1	2	51:49
1	4	36:64
1	8	52:48
1	16	0:100

Note: Cost of IN=IB=1 is the cost used to obtain the rules presented in Appendix II

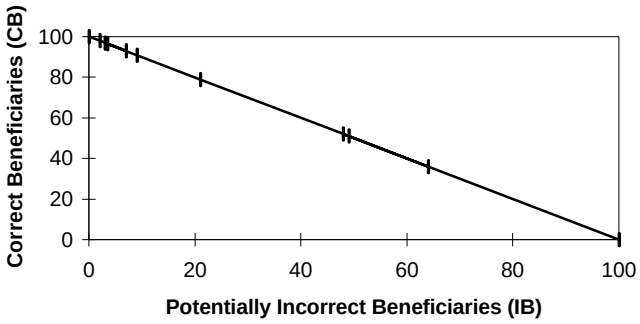
Table VII Effect of increasing cost of error IB greater than IN

Cost of error IB	Cost of error IN	Proportion CB:IB
1	**1**	**79:21**
2	1	91:9
4	1	97:3
8	1	96.5:3.5
16	1	93:7
32	1	98:2
64	1	100:0

Note: Cost of IN=IB=1 is the cost used to obtain the rules presented in Appendix II

As the cost of increasing error IN increases to higher than that of error IB, the proportion of potentially incorrect beneficiaries identified increases and the correct beneficiaries decreases and vice versa (see Figure II).

Figure II Change in proportion of CB Vs IB as cost of IN increases



It is obvious from Tables VI and VII as well as Figure II that, as making one type of error over another becomes increasingly costly, at extreme cost 100% of the cases are identified in the group which minimises the cost of that error. Thus the classification

rules would be extremely generalised so that either all households are classified as 'Poor' or as 'Non-Poor'. Caution would therefore have to be exercised in policy applications when using very high costs for either error.

5.3 Contributions and shortcomings

Contributions and shortcomings specific to our study (rather than the analysis method used) are the following. First, our data were census data. This had the advantage of allowing an assessment of a large number of variables without pre-selection based on either theoretical or practical considerations. Their disadvantage is that some variables that might have been included if the data were collected specifically with poverty profiles in mind may have been excluded. Some examples are accidents and episodes of acute sickness to members of the household, and to draught animals; details of sanitation, alcohol consumption and access to protected drinking water. Second, as discussed earlier, a number of variables included in the analysis and identified as bearing an important relationship to income are in themselves susceptible to fudging. If data were collected for the specific purpose of investigating the issue explored here, they would have to be restricted to variables that are transparent and hard to manipulate. Finally, while the method of analysis itself is inexpensive (requiring the purchase of particular soft-ware) and provides clear and easily interpretable results, any policy application of such an analysis directed to this particular end would have to take into account the cost in time and resources required to: a) collect variables for analysis and b) follow-up and confirm if households identified as potentially incorrect beneficiaries are indeed not eligible for the receipt of benefits. Political, financial, social and other factors will bear on the relative desirability of identifying and investigating potentially incorrect beneficiaries at the cost of reducing the number of households considered eligible for the direct receipt of benefits.

With regard to the methodology, we have introduced here a method of multivariate analysis which offers some advantages over traditional regression based methods. The class probability tree analysis takes into account non-linear relationships and the results are expressed as easily understandable rules. Some potential uses are as follows:

1. The class-probability analysis may also be useful to explore relationships between income (or health or education or nutrition indicators) and other variables with a view to identifying appropriate proxy variables.
2. The results can also be used to give an insight into the features associated with, or responsible for, the dimension of poverty being explored.
3. The analysis can help identify households or individuals that have been wrongly included in the target group. Our analysis, for example, identifies variables useful to further assess households, to confirm their eligibility for the receipt of state transfers.

The presentation of the results as easily interpretable rules make this analysis method particularly attractive for use by those involved in policy formulation and implementation.

References

Agnihotri,. "**Sex Ratio Imbalances in India - A Disaggregated Analysis**", PhD. Thesis, University of East Anglia, Norwich, 1997.

Ardener, S. "Women Making Money Go Round: ROSCAs Revisited" in Ardener, S. and Burman, S. (eds.) **Money-Go-Rounds The importance of Rotating Savings and Credit Associations for Women**, Berg Publishers Limited, Oxford/Washington DC, 1995.

Baulch, B. and McCulloch, N. **Being Poor and Becoming Poor: Poverty, Status and Poverty Transitions in Rural Pakistan**, Poverty Research Programme Working Paper 79, Institute of Development Studies, Sussex University, Brighton, 1998.

Behrman, J. and Srinivasan, T.N.(eds.) **Handbook of Development Economics**, Vol. 3, Elsevier Science B.V., Netherlands, 1995.

Bliven, N., Ramasamy, C. and Wanmali, S. "**Devising Policies to help the poor in South India**", International Food Policy Research Institute, Washington, D.C., 1997.

Calomiris, C.W. and Rajaraman, I. "The role of ROSCAs: lumpy durables or event insurance?", **Journal of Development Economics** 56, 207-216, 1998

Clark, P. and Niblett, T. "The CN2 Algorithm", **Machine Learning**, 3:262-283, 1989.

Colatei, D. and Harriss-White., "The Classification of Rural Households", chapter in Harriss-White, B. (ed.) **Adjustment and Development in South India**, Sage, New Delhi, forthcoming.

Copestake, J., "The Integrated Rural Development Programme: Performance during the Sixth Plan, Policy Responses and Proposals for Reform" in Harriss, B., Guhan, S. and Cassen, R.H. (eds.),op. cit., 1992.

Cornia, G.A. and Stewart, F. "Food Subsidies: Two Errors of Targeting" in Stewart (ed.) **Adjustment and Poverty**, Routledge, London and New York, 1995.

Cornia, G.A., Jolly, R. and Stewart, F. (eds.) **Adjustment with a Human Face**, UNICEF, Clarendon Press, Oxford, 1987.

Dandekar, V.M. and Rath, N. (1971). **Poverty in India**, Indian School of Political Economy, Poona.

Dreze, J., Lanjouw, P. and Sharma, N. (1998) "Economic Development in Palanpur, 1957-93", pp 114-234 in Lanjouw, P and Stern, N. (eds.), op. cit., 1998.

Dreze, J., and Sharma, N. (1998) "Palanpur: Population, Society, Economy", pp 3-113 in Lanjouw, P and Stern, N. (eds.), op. cit., 1998.

Erb, S. and Harriss-White. "Outcast from the Social Welfare Agenda: A Study of the Economic Penalties of Disability in Rural Tamil Nadu", in Harriss-White, B. (ed.) **Adjustment and Development in South India**, Sage, New Delhi, forthcoming.

Farmer, B.H., Madduma Bandara, C.M., Shanmuga Sundaram, V. and Silva, W.P.T. "Setting the Stage" in Farmer, B.H. (ed.) **Green Revolution? Technology and Change in Rice-growing Areas of Tamil Nadu and Sri Lanka**, Macmillan, London and Basingstoke, 1977.

Gaiha, R. (1988) "On Measuring the Risk of Rural Poverty in India" in Srinivasan, T.N. and Bardhan, P.K. (eds.), op. cit., 219-261.

Glewwe, P. and Kanaan, O. "Targeting Assistance to the Poor: A multivariate Approach Using Household Survey Data", **Development Economics Research Centre**, Discussion Paper 94, Warwick University, U.K, 1989.

Greer, J. and Thorbecke, E. "A Methodology for Measuring Food Poverty Applied to Kenya", **Journal of Development Economics**, 24, 59-74, 1986.

Guhan, S. and Mencher, J.P. "Iruvelpattu Revisited", Economic and Political Weekly, 18 (23 and 24), 1983.

Hanmer, L., Pyatt, G. and White, H. (1997) **Poverty in Sub-Saharan Africa - What can we learn from the World Bank's Poverty Assessments**, Institute of Social Studies, The Hague, The Netherlands.

Harriss, B. "Rural Poverty in India: Micro-level evidence" in Harriss, B., Guhan, S. and Cassen, R.H. (eds.),op. cit., 1992.

Harriss, B., Guhan, S. and Cassen, R.H. (eds.), **Poverty in India: Research and Policy**, Oxford University Press, New Delhi, 1992.

Harriss-White, B. "Economic Restructuring: State, Market, Collective and Household Action in India's Social Sector", **European Journal of Development Research**, 7(1), 1995.

Harriss-White, B. and Janakarajan, S. "From Green Revolution to Rural Industrial Revolution", **Economic and Political Weekly**, 32 (25), 1997.

Harriss-White, B. and Subramanian, S. (eds.) **Illfare in India: Essays on India's Social Sector in Honour of S.Guhan**, Sage, New Delhi and London, 1999.

Hazell, P. B. R. and Ramasamy, C. **The Green Revolution Reconsidered: The Impact of High-Yielding Rice Varieties in South India**, Johns Hopkins, Baltimore, 1991.

Janakarajan, S. "Complexities of Agrarian Markets and Agrarian Relations: A Study of Villages in Northern Tamil Nadu", draft for Dissemination Workshop for the ODA Funded Project on **Adjustment and Development: Agrarian Change, Markets and Social Welfare in South India 1973-1993,** MIDS, Madras, 1996.

Jayaraj, D. and Subramanian, S. "Poverty and Discrimination: Measurement and Evidence from Rural India" in Harriss-White, B and Subramanian, S.(eds.) op. cit., 1999, p 196-224.

Jayaraj, D. "Determinants of Rural Non-agricultural Employment", in Subramanian, S.(ed.), op. cit., 1992.

Jayaraj, D. "Structural transformation of the Rural Workforce in Tamil Nadu: An Analysis of the Impact of Social Factors", draft for Dissemination Workshop for the ODA Funded Project on **Adjustment and Development: Agrarian Change, Markets and Social Welfare in South India 1973-1993,** MIDS**,** Madras, 1996.

Kohavi, R. and John, G.H. "Automatic Parameter Selection by Minimizing Estimated Error", in Prieditis, A. and Russell, S. (eds**.) Machine Learning: Proceedings of the Twelfth International Conference**, Morgan Kaufmann Publishers, San Francisco, CA, 1995

Krishnaji, N. "The Size and Structure of Agricultural Labour Households" in Rodgers (ed.), op. cit., 1989, p 121-150.

Lanjouw, P and Stern, N. eds. (1998) Economic Development in Palanpur over Five Decades, Clarendon Press, Oxford.

Lipton, M. and Maxwell, S. **The New Poverty Agenda: an Overview**, Discussion Paper No. 306, Institute of Development Studies, 1992.

Lipton, M. and Ravallion, M. "Poverty and Policy" Chapter 41 in Behrman, J. and Srinivasan, T.N., (eds.), op. cit., 1995.

Maxwell, S. (1999). "The Meaning and Measurement of Poverty", **ODI Poverty Briefing Series**, Overseas Development Institute, London. Also available on the World Wide Web at http://www.oneworld.org/odi/briefing/pov3.html

Michie, D. **Quality Control of Induced Rule Based Programs: The Fifth Generation**, The GS Institute, London, 1984.

Michie, D., "Personal models of rationality", **Journal of Statistical Planning and Inference**, 25: 381-399, 1988.

Paul, S. "A Model of Constructing the Poverty Line" **Journal of Development Economics**, 30, 129-44, 1989.

Planning Commission, Government of India, **Report of the Expert Group on Estimation of Proportion and Number of Poor**, 1993.

Quinlan, J.R. **C4.5 Programs for Machine Learning**. Morgan Kaufmann, San mateo CA, 1993.

Quinlan, J.R. **See5: An Informal Tutorial**, notes accompanying the See5 software; http://www.rulequest.com/see5-win.html, 1998.

Ramu, G.N. Family Structure and Fertility, Sage, New Delhi and London, 1988.

Ravallion, M. and Bidani, B. "How Robust is a Poverty Profile?" **World Bank Economic Review**, 8, 75-102, 1994.

Ravallion, M. and Sen, B. "When Method Matters; Toward a Resolution of the Debate about Bangladesh's Poverty measures", **Economic Development and Change**, 44, pp761-92, 1996.

Rodgers, G. (ed.) **Population Growth and Poverty in Rural South Asia**, Sage, New Delhi, 1989.

Ruggeri Laderchi, C (1997) "Poverty and its many dimensions: the role of income as an indicator", **Oxford Development Studies**, 25 (3), 345-360.

Saith, R.R. "The Practical Identification of Poor Households: a Decision Tree Analysis of Village Census Data in Northern Tamil Nadu: Preliminary Results" **Adjustment and Development Project**, Working paper Number 38, Queen Elizabeth House, Oxford, UK, 1996

Saith, R.R., Srinivasan, A, Michie, D. and Sargent, I.L. "Relationships between the developmental potential of human in-vitro fertilization embryos and the features describing the embryo, oocyte and follicle", **Human Reproduction Update**, 4 (1), 1998.

Saith, R.R. and Harriss-White, B. "The Gender Sensitivity of Well-Being Indicators", **Development and Change**, 30, pp 465-497, 1999.

Sanyal, S.K. "Trends in Landholding and Poverty in Rural India" in Srinivasan, T.N. and Bardhan, P.K. (eds.), op. cit., 1988, p 121-153.

Singh, K.S. **People of India**, National Series, Volume II, Anthropological Survey of India, OUP, Delhi, Bombay, Calcutta, Delhi, 2nd revised edition, 1995.

Srinivasan, T.N. and Bardhan, P.K. (eds.) **Rural Poverty in South Asia**, OUP, New Delhi, 1988.

Subramanian, S. (ed**.) Themes in Development Economics: Essays in Honour of Malcolm Adiseshiah**, OUP, New Delhi, 1992.

Subramanian, S. and Harriss-White, B. "Introduction" in Harriss-White, B and Subramanian, S.(eds.) op. cit., 1999, p 17-43.

Weiss, S.M. and Kulikowski, C.A. **Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems**, Morgan Kaufmann, San Mateo, CA, 1991.

Wodon, Q. "Food Energy Intake and Cost of Basic Needs: Measuring Poverty in Bangladesh", **The Journal of Development Studies**, 34 (2), 1997a.

Wodon, Q. "Targeting the poor using ROC curves", **World Development**, 25 (12), 1997b.

6. Appendices

6.1 Appendix I

The Table below lists the 478 variables used in the analysis. Unless mentioned otherwise, each variable is related to the household as a whole, rather than any particular individual within the household. The relationship between these variables and household income was analysed using the technique of class probability tree analysis.

273 of the variables used in the analysis were recorded in Census data. 205 were however constructed specifically for inclusion in the analysis using information available in the census data. A mention of whether the variable was used as recorded (R) or was constructed (C) is made in the Table below. The last column also contains information on whether the variable was included in the analysis restricted to non-fudgable (NF) variables

Table of features

Nos.	Variable	R/ C	NF
1	Caste: Udayar; Naidu; Yadavas; Naicker, Nadar; Scheduled castes; Adhidravidars; Parayar; A.Mudaliar; S.Vellalur; Chettiyar; Brahmins; Devadasi; Viswakarna (gold, blacksmith) and Aachary; Vanniars; Pandidar and Navthar; Vettaikaran; Christian (scheduled caste); Pandaram (poojali); Upparavar (chettiyar); Kuzavar (potter); Karuneeyar; Boer; Sengunthar; Dhobi; Muslim; Reddiyar; Gramini; Oddar; Arunthathiyar; or Nainar	R	
2	Scheduled status of caste: Yes or no	C[29]	NF
3 - 12	Nos. of male members belonging to the following age groups: [30]0 - 1, 0 - 4, 0 - 9, 0 -14, 10 - 14, 15 -34, 15 - 64 (working members), 35 - 49, ≥ 65 (elderly), < 15 - > 64 (dependant members)	C	NF
13 - 22	Nos. of female members belonging to the following age groups: 0 - 1, 0 - 4, 0 - 9, 0 -14, 10 - 14, 15 -34, 15 - 64, 35 - 49, ≥ 65, < 15 - > 64	C	NF
23	Total dependants (male + female members)	C	NF
24	Total working age population (male + female members)	C	NF
25	Dependency ratio (i.e. variable 22/ variable 23)	C	NF
26	Gender of head of household[31]	C	NF

[29] This variable was constructed as follows. Households classified as belonging to a Scheduled caste as well as households belonging to the caste groups Adhidravidars, Parayar and Arunthathiyar were included in the scheduled caste group (following the classification suggested in the Anthropological Survey of India by Singh, 199%). All other households were labelled non-scheduled caste.

[30] Individuals were grouped under these age categories as these groups have been identified in the literature as showing significant differences in mortality rates, earning potential and/ or fertility rates. The number of household members belonging to each of these age groups, could thus be hypothesised as bearing an important relationship with household income.

[31] The head of the household was considered to be female in the following households: (a) all households with only female members; (b) households with male and female members, but with only

27 - 35	Nos. of male members with the following levels of education: illiterate; primary (up to class seven); middle (up to class nine); secondary or matriculation (class 10); higher secondary or pre-university; diploma; literate with formal education (less than class five); graduates and above; or too young to be in school (usually between 0 - 4 years)	C	
36 - 44	Nos. of female members with one of the following levels of education: illiterate; primary (up to class seven); middle (up to class nine); secondary or matriculation (class 10); higher secondary or pre-university; diploma; literate with formal education (less than class five); graduates and above; or too young to be in school (usually between 0 - 4 years)	C	
45 - 73	Nos. of male members whose primary occupation in the past belonged to one of the following categories: inactive or no occupation; too young (usually pre-school i.e. between 0-4 years); inactive due to handicap or illness; student; cultivation (either own or tenancy); animal husbandry; agricultural labourer; weaver (own); weaver (coolie); weaving assistant at home; weaving assistant and apprentice coolie; manufacturing (household industry other than weaving e.g. carpenter, basketmaking, sekkaduthal i.e. oil extraction); manufacturing (other than household industry e.g. industrial worker, factory worker, centering works, brick making, beedi); construction e.g. pipeline laying, mason; trade and commerce (including fishing); transport, storage and communication (including van operator, driver, lorry leader, auto) service sector e.g. teacher, barber, watchman, company employee, bank cashier, tailor, lift mechanic, cook, registered medical practitioner, medical assistant, dhobi, military services, gardener, maniakaran (village headman), village assistant officer, karumnar (black smith); housework and baby-sitting; fire wood collection, noon meal scheme organiser or cook, cattle broker and commission agent; padiyal (labourer under a type of permanent labour system); kootali (a different type of permanent labour system); tailor; tree climbing (or pambai); karnam (village accountant); kammukutti (community irrigation worker) and thalayari i.e. village assistant; toddy tapper; or miscellaneous e.g. band player, maternity helper, fishing, non-agricultural labourer, dancer, composer, earth -works, pot-works, electric-works, stone-works, leather works, iron-works, bore-well works, electricity board wiring; painting, mechanic apprentice, courier, lorry cleaner, lorry loader, saloon worker, canteen staff, motor winding and tailoring apprentice, astrologist, thatcher, community worker organiser, pensioner, recipient of a form of Government welfare programme called TUNIP and sekkaduthal.	C	
74	Total nos. of male members who had a primary occupation in the past	C	
75 - 103	Nos. of male members who have a primary occupation at present, in any one of the following categories: same categories as for 43 - 71	C	
104	Total nos. of male members who have a primary occupation at present	C	
105	Name of .village household belongs to: Vegamangalam, Srirungathur, Thuli, Vengodu, Vayalur, Mepathurai, Amudhur or Kalpathu	R	NF
106	Total nos. of male members who have a secondary occupation	C	
107 - 135	Nos. of female members who in the past had a primary occupation in the following categories: same categories as for 43 - 71	C	
136	Total nos. of female members who had a primary occupation in the past (including housework as a occupation)	C	
137	Total nos. of female members who had a primary occupation in the past (excluding housework as a occupation)	C	
138 - 166	Nos. of female members having a primary occupation at present, in the following categories: same categories as for 43 - 71	C	
167	Total nos. of female members who have a primary occupation at present (including housework)	C	
168	Total nos. of female members who have a primary occupation at present (excluding housework)	C	

females in the working age group (15-64). The head of the household was assumed to be male in all the other households.

169	Total nos. of female members who have a secondary occupation (including housework)	C	
170	Total nos. of female members who have a secondary occupation (excluding housework)	C	
171	Total nos. of male members	R	NF
172	Total nos. of female members	R	NF
173	Total nos. of members (variables 169 + 170)	C	NF
174	Nos. of male members who are from same village	C	
175	Nos. of male members who have come in from outside the village	C	
176	Nos. of female members who are from same village	C	
177	Nos. of female members who have come in from outside the village	C	
178	Nos. of male members who are recipients of any of the government social welfare schemes.[32]	C	
179	Nos. of female members who are recipients of any of the government social welfare schemes	C	
180	Nos. of ill or handicapped male members	C	NF
181	Nos. of ill or handicapped female members	C	NF
182	Type of family - nuclear or joint	R	NF
183	Total working male members[33]	R	
184	Total working female members	R	
185	Nos. of members who migrated out[34] - one or more than one	C	
186	Nos. of members who migrated in - one or more than one	C	
187 - 188	Year and month of outmigration of member 1[35]	R	
189	Name of place outmigrated from	R	
190	Present occupation of outmigrated member - same categories as for 43 - 71	R	
191	Nos. of months member outmigrated for	R	
192	Nos. of outmigrants who are presently employed (self or otherwise) - one or more than one[36]	C	

[32] The government social welfare schemes belong to the following categories: old age pension; maternity benefits; widow benefit; integrated rural development programme (IRDP); public distribution system (PDS) or free sari or free dhoti; noon-meals scheme, dhara scheme - well digging; free house; TUNIP - a form of old age pension scheme; and others.

[33] The criterion used by the census data recorders to decide on whether a member was considered a working member (e.g. inclusion of part-time work or only full-time work) is not clear. The values entered under this variable differ from those calculated for the variables related to the total number of males or females that currently have a primary occupation.

[34] Information on this variable has been recorded in the census data, for at most two members of each household. For households with information on just one member, it is clear that only one member migrated. If information is recorded for two members however, we cannot know for sure whether or not more than two members have migrated.

[35] Since most of the houses that had any migration (either in or out), had only one member migrating, we decided to include in the analysis only information recorded for one member as a representative of the household. In cases with information for two members, the first member for whom information was recorded was chosen although as far as we are aware the order of recording was arbitrary.

193	Name of place lace from which member has inmigrated	R
194	Year of inmigration	R
195	Reason for inmigration	R
196	Past occupation of inmigrant - same categories as for 43 - 71	R
197	Nos. of inmigrants who were employed (self or otherwise) in the past - one or more than one[37]	C
198 - 203	Extent of own land of following categories inside village (acres)[38] Nanjai Sole Surface (NSS) land Nanjai surface + well (NSW) land Punjai with well (PWW) land Punjai rain fed (PRF) land Other Total (NSS + NSW + PWW+ PRF + Other)	R
204 - 208	Value of own land of following categories inside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW+ PRF)	all R except total which was C
209 - 214	Extent of own land of following categories outside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
215 - 219	Value of own land of following categories outside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
220 - 223	Value of total (inside + outside village) own land of following categories: NSS; NSW; PWW; PRF	C
224	Extent of total (inside + outside village, NSS + NSW + PWW + PRF + Other) own land	C
225	Value of total (inside + outside village, NSS + NSW + PWW + PRF + Other) own land	C
226 - 231	Extent of leased in land of following categories inside village (acres): NSS; NSW; PWW; PRW; Other; Total	R
232 - 236	Rent paid for leased in land of following categories inside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
237 - 242	Extent of leased in land of following categories outside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
243 - 247	Rent paid for leased in land of following categories outside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
248 - 251	Rent paid for total (inside + outside village) leased in land of following categories: NSS; NSW; PWW; PRF	C

[36] Details on migration have been recorded for at most two members of each household. For households with information on migration and occupation for two members, we cannot know for sure whether or not more than two members have migrated.

[37] Details on migration have been recorded for at most two members of each household. For households with information on migration and occupation for two members, we cannot know for sure whether or not more than two members have migrated.

[38] This is a old classification used for land revenue purposes by the British, based on the extent and type of irrigation. Nanjai was referred to as wet land and was land irrigated by tank (open reservoir) water. Remaining land was referred to as Punjai or dry land. With the advent of well based irrigation however, dry land receiving adequate quantities of water too could be considered wet. Further, with the passage of time, wetland has been increasingly less reliant on tank water and more dependant on well irrigation due to the depletion of the water table caused by well irrigation. With these changes therefore, land values and productivity for the different types of land converge and the classification is less meaningful.

252	Extent of total (inside + outside village, NSS + NSW + PWW + PRF + Other) leased in land	C
253 - 258	Extent of leased out land of following categories inside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
259 - 263	Leased out land of following categories inside village value (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
264 - 268	Rent received for land of following categories leased out inside village: NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
269 - 274	Extent of leased out land of following categories outside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
275 - 279	Value of land of following categories leased out outside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
280 - 284	Rent received for land of following categories leased out outside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	R
285 - 288	Value of total (inside + outside village) land of following categories leased out: NSS; NSW; PWW; PRF	C
289	Value of total leased out (inside + outside village, NSS + NSW + PWW + PRF) land	C
290 - 293	Total rent received for leased out (inside + outside village) land of following categories: NSS; NSW; PWW; PRF	C
294	Extent of total (inside + outside village, NSS + NSW + PWW + PRF + Other) leased out land	C
295 - 300	Extent of land of following categories mortgaged in inside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
301 - 305	Mortgage value of land of following categories mortgaged in inside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
306 - 311	Extent of land of following categories mortgaged in outside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
312 - 316	Mortgage value of land of following categories mortgaged in outside village: NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW +PRF)	all R except total which was C
317 - 320	Mortgage value of total (inside + outside village) land of following categories mortgaged in: NSS; NSW; PWW; PRF	C
321	Extent of total (inside + outside village, NSS + NSW + PWW + PRF + Other) land mortgaged in	C
322 - 327	Extent of land of following categories mortgaged out inside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
328 - 332	Value of land of following categories mortgaged out inside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
333 - 337	Mortgage value of land of following categories mortgaged out inside village NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
338 - 343	Extent of land of following categories mortgaged out outside village (acres): NSS; NSW; PWW; PRF; Other; Total	R
344 - 348	Value of land of following categories mortgaged out outside village (rupees) NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
349 - 353	Mortgage value of land of following categories mortgaged out outside village (rupees): NSS; NSW; PWW; PRF; Total (NSS + NSW + PWW + PRF)	all R except total which was C
354 - 357	Value of total land mortgaged of following categories out (inside + outside village): NSS; NSW; PWW; PRF	C
358	Value of total land mortgaged out (inside + outside village, NSS +NSW + PWW + PRF)	C
359 - 362	Mortgage value of total land of following categories mortgaged out (inside + outside	C

	village): NSS; NSW; PWW; PRF		
363	Extent of total (inside + outside village, NSS + NSW + PWW + PRF + Other) land mortgaged out	C	
364 - 369	Extent of operational holding of following categories inside village (acres): NSS; NSW; PWW; PRF; Other; Total	R	
370 - 375	Extent of operational holding of following categories outside village (acres): NSS; NSW; PWW; PRF; Other; Total	R	
376 - 380	Extent of total (inside + outside village) operational holding of following categories: NSS; NSW; PWW; PRF; Other	C	
381	Extent of total (inside + outside village, NSS + NSW + PWW + PRF + Other) operational holding	C	
382	Net rent (rent paid for total land leased in + rent received for total land leased out)	C	
383	Net mortgage (total value of mortgage received for land mortgaged out + total value of mortgage paid for land mortgaged in)	C	
384 - 386	Total area of irrigated land on which crops grown (acres)in: Season 1[39]; Season 2; Season 3	R	
387 - 389	Total area of un-irrigated land on which crops grown (acres) in: Season 1;Season 2; Season 3	R	
390	Gross irrigated area on which crops grown (Season 1 + 2 + 3)	R	
391	Gross un-irrigated area on which crops grown (Season 1 + 2 + 3)	R	
392 - 394	Production from irrigated land (KGs) in: Season 1; Season 2; Season 3	C	
395 - 397	Production over un-irrigated land (KGs) in: Season 1; Season 2; Season 3	C	
398 - 400	Total production (irrigated + un-irrigated) in: Season 1; Season 2; Season 3	C	
401	Gross production from irrigated land (season 1 + season 2 + season 3)	C	
402	Gross production from un-irrigated land (season 1 + season 2 + season 3)	C	
403	Gross overall production (399 + 400)	C	
404 - 409	Total (irrigated + un-irrigated) land area over which following crops grown: paddy; groundnut; sugarcane; ragi; plantain; other crops [40]	C	
410	Nos. of wells	R	NF
411	Estimated monetary value of wells (rupees)	R	NF
412	Estimated monetary value of electrical pump sets (rupees)	R	NF
413	Estimated monetary value of oil engines (rupees)	R	NF
414	Estimated monetary value of tractors (rupees)	R	NF
415	Estimated monetary value of power tills (rupees)	R	NF

[39] Season 1 indicates the Samba season extending from Aug-Sept to Dec-Jan; Season 2 indicates the Navarai season extending from Dec-Jan to April-May; Season 3 indicates the Sornawari season extending from May-June to July-August.

[40] The group other crops includes the following crops: gingelly, mulberry, chilli, mango, cotton, root vegetable, corn and onion, tomato, brinjal, other vegetables, turmeric, kara, cholam, ulundu and green gram.

416	Estimated monetary value of sprays (rupees)	R	NF
417	Estimated monetary value of traditional bullock carts (rupees)	R	NF
418	Estimated monetary value of modern bullock carts (rupees)	R	NF
419	Estimated monetary value of agricultural implements (rupees)	R	NF
420	Estimated monetary value of plough bullocks (rupees)	R	
421	Estimated monetary value of cart bullocks (rupees)	R	
422	Estimated monetary value of milch animals (rupees)	R	
423	Estimated monetary value of sheep and goat (rupees)	R	
424	Estimated monetary value of other agricultural assets possessed (rupees)	R	
425	Total value of agricultural assets possessed (rupees)	C	
426 - 427	Extent of homestead land (acres): inside village outside village	R	 NF
428 - 429	Estimated monetary value of homestead land (rupees): inside village outside village	R	 NF
430 - 431	Nos. of buildings inside village outside village	R	 NF
432 - 433	Estimated monetary value of buildings (rupees): inside village outside village	R	 NF
434 - 435	Estimated monetary value of business assets (rupees): inside village outside village	R	 NF
436	Weight of jewellery possessed	R	
437	Value of jewellery possessed (rupees)	R	
438	Cash balance possessed (rupees)	R	
439	Total monetary value (rupees) of non- agricultural assets possessed (including cash balance)	C	
440	Institutional liabilities	R	
441 - 447	Non-institutional liabilities owed to the following: money lenders/pawnbrokers; landlords; traders/commission mandis; silk maligai; friends/relatives; others; total (including all the above)	R	
448	Total liabilities (institutional + total non-institutional)	C	
449 - 555	Details of chit fund held by member 1[41] total amount; nos. of instalments; payment frequency; realised/not realised; bidded amount; if realised nos. of instalments due; if realised how was it spent	R	

[41] Data were available for up to three members of the household. However while the number of households with at least one member with chit was approximately 261, only about 60 households had a second member and only 20 households had a third member with a chit fund. We thus decided to include only information related to the first member as a representative of household. Any influence of more than one member having a chit fund would be taken into account by variables 454 - 456 which relate to the full household.

456	Total nos. of members in household who have a chit fund	C	
457	Total amount of chit (including all members of household who have a chit fund)	C	
458	Average amount of chit	C	
459	Water: purchased; sold ; neither	R	
460	Water purchased last year: yes; no	C	
461	Water sold last year: yes, no	C	
462 - 463	Water purchased last year for following area of land (acres): Gross wet area Gross dry area	C	
464 - 465	Water sold last year for following area of land (acres): Gross wet area Gross dry area	C	
466 - 468	Payment for water purchased: nos. of bags of paddy nos. of bags of groundnut cash (rupees)	C	
469 - 471	payment received for water sold: No. of bags of paddy No. of bags of groundnut cash (rupees)	C	
472	Condition of house: own; rented; shared; poromboke (government land); group house; belonging to relatives or other friends; a mixture of any two of the above categories; no house	R	
473	Nos. of occupants: single household or if multiple, no. of households	C	NF
474	Roof of house: thatched; tiled; terraced; a mixture of any two of the above categories; no roof (as no house)	R	NF
475	Wall of house: brick; mud; stone; stone (probably different kind of stone); thatched; a mixture of any two of the above categories; no wall (as no house)	R	NF
476	Floor of house: mud; cement; mosaic; a mixture of any two of the above categories; no floor (as no house)	R	NF
477	Connected to a source of electricity: yes or no	R	NF
478	Medium of cooking: gas; kerosene; firewood; a mixture of any two of the above categories; no medium for cooking (as no house)	R	NF

6.2 Appendix II

The Rule Set of 26 rules that was obtained by analysing data for 1500 households in the training data, with 478 variables for each household, identified some 40 variables as bearing an important relationship with the income of the household. The predictive error of the Rule Set is represented by the error estimates shown in Tables IV and V in the Results section.[42]

The 6 rules predicting households as 'Poor' are presented first followed by the 10 rules predicting households as 'Non-poor'. Two values associated with each rule are presented viz. (A) the number of training cases covered by each rule and (B) the accuracy of the rule i.e. probability of a covered case belonging to the class predicted by the rule. Rules have been presented based on the A*B with the highest value for each class being presented first. This is followed by the default rule[43].

Rules predicting class 'Poor'.

Rule 1 (covers 467 cases)

[42] The Rule Set that gave the error estimates shown in Tables IV and V had 27 rules. The largest number of cases covered by any single rule was 511 and the smallest number was 4. Since the rules that classify a small number of cases would be expected to contribute little to the overall error rate, an investigation was undertaken to see if progressive exclusion of the rules covering a small number of cases caused any significant changes in the estimated error values and the proportion of CB:IB cases. A calculation of these values as each Rule (beginning with the rule covering the smallest number of cases and then progressively increasing) was dropped showed that dropping up to 10 rules did not alter the overall error or distribution of cases significantly (the test set error was found to be 25 and the CB:IB proportion 81.0% :19.0% as compared to test set error rates of 24% and CB:IB proportion of 78:22 for the complete 27 Rule Set). The remaining 17 Rules of the Rule Set are presented here.

[43] When constructing rules from the tree, See5 excludes some rules that are superfluous or whose exclusion is not expected to lower the predictive accuracy. This may leave some cases unclassified by any rule. To cover for this, the conversion to rules always introduces a 'default' rule, which specifies how such cases are to be classified.

If

no male member had a miscellaneous primary occupation in the past,[44]
no male member is currently primarily occupied in construction work,[45]
no male member is currently primarily occupied in the service sector,[46]
the value of total Punjai Rain Fed (PRF) land owned is less than or equal to Rs. 32500,
value of total land leased out is at most Rs.25000,
total area of irrigated land over which crops were grown in Season 1 is at most 1.5 acres,
gross overall production of all crops is at most 1750 KGs,
total area of land over which plantains were grown is less than or equal to 0.33 acres,
does not own any oil engines,
business assets in village have an estimated monetary value of at most Rs. 1500,
total amount of chit fund is at most Rs. 15000 and
total nos. of members is more than three

Then

Probability of 'Poor' is 84% and 'Non-poor' is 16% (accuracy = 0.84)

Class prediction is taken as **Poor**

Rule 2 (covers 511cases

If

at least one (or more) female members is currently an agricultural labourer,
value of total land leased out is at most Rs.25000,
gross overall production of all crops is at most 1750 KGs and
total amount of chit fund is at most Rs. 15000

Then

Probability of 'Poor' is 77% and 'Non-poor' is 23% (Accuracy = 0.77)

Class prediction is taken as **Poor**

Rule 3 (covers 207 cases)

---

[44] Miscellaneous occupations include a wide variety of occupations that do not fall into the other categories. Examples are: band player, maternity helper, fishing, non-agricultural labourer, dancer, composer, earth -works, pot-works, electric-works, stone-works, leather works, iron-works, bore-well works, electricity board wiring; painting, mechanic apprentice, courier, lorry cleaner, lorry loader, saloon worker, canteen staff, motor winding and tailoring apprentice, astrologist, thatcher, community worker organiser, pensioner, recipient of a form of Government welfare programme called TUNIP and sekkaduthal (i.e. oil extraction).

[45] Construction work also includes masons and pipeline work

[46] Service sector includes a range of occupations like: teacher, barber, watchman, company employee, bank cashier, tailor, lift mechanic, cook, registered medical practitioner, medical assistant, dhobi, military services, gardener, maniakaran (village headman), village assistant officer, karumnar (black smith); housework and baby-sitting; fire wood collection, noon meal scheme organiser or cook, cattle broker, commission agent and tailor (sometimes tailor was included in this category although more regularly tailor is considered as a separate occupation category)

If
at least one (or more) female member is working,[47]
gross overall production of all crops is at most 1750 KGs and
total amount of chit fund is at most Rs. 15000
Then
Probability of 'Poor' is 77% and 'Non-poor' is 23% (Accuracy = 0.77)

Class prediction is taken as **Poor**

Rule 4 (covers 136 cases)

If
no male member currently has a primary occupation,
the value of total Punjai Rain Fed (PRF) land owned is at most Rs. 32500,
value of total land leased out is at most Rs.25000,
gross overall production of all crops is at most 1750 KGs,
total area of land over which plantains were grown is less than or equal to 0.33 acres and
total amount of chit fund is at most Rs. 15000
Then
Probability of 'Poor' is 86% and 'Non-poor' is 14% (Accuracy = 0.86)
Then

Class prediction taken as **Poor**

Rule 5: (covers 113 cases)

If
at least two (or more) male members are currently students,
gross overall production of all crops is at most 1750 KGs,
total area of land over which plantains were grown is less than or equal to 0.33 acres and
business assets in village have an estimated monetary value of at most Rs. 1500
Then
Probability of 'Poor' is 81% and 'Non-poor' is 19% (Accuracy = 0.81)

Class prediction is taken as **Poor**

Rule 6 (covers 60 cases)

If
at least one (or more) female member is aged between 0 and 1 years and
gross overall production of all crops is at most 1750 KGs
Then
Probability of 'Poor' is 73% and 'Non-poor' is 27% (Accuracy = 0.73)

Class prediction is taken as **Poor**

[47] The total number of working males and females was recorded in the data. It is not clear however how this figure was obtained. The values differ from the variable total number of males (or females) that currently have a primary occupation which has been constructed from data available in the census.

Rule 7 (covers 12 cases)

If
at least one (or more) male members are currently primarily occupied in the service sector and
the estimated monetary value of business assets in the village is greater than Rs 0 but at most Rs 1500
Then
Probability of 'Poor' is 79% and 'Non-poor' is 21% (Accuracy = 0..79)

Class prediction is taken as **Poor**

Rule 8 (covers 11 cases)

If
at least one (or more) male members is currently an agricultural labourer,
total area of land owned is at most 1.62 acres and
gross overall production of all crops is greater than 1750 KGs
Then
Probability of 'Poor' is 85% and 'Non-poor' is 15% (Accuracy = 0.85)

Class prediction is taken as **Poor**

Rule 9 (covers 6 cases)

If
the number of male dependent members (aged under 15 or over 64) is greater than 2,
the total extent of operational holding land in the village is at most 1.75 acres and
gross overall production of all crops is greater than 1750 KGs
Then
Probability of 'Poor' is 88% and 'Non-poor' is 12% (Accuracy = 0.88)

Class prediction is taken as **Poor**

Rule 10 (covers 4 cases)

If
at least one (or more) male member had a miscellaneous primary occupation in the past,
at least one (or more) male members is currently an agricultural labourer and
the estimated monetary value of plough bullocks owned is at most Rs 2000
Then
Probability of 'Poor' is 83% and 'Non-poor' is 17% (Accuracy = 0.83)

Class prediction is taken as **Poor**

Rules predicting class 'Non-poor'

Rule 1 (covers 328 cases)

If
at most two male members are dependant,
no male member is currently an agricultural labourer and
gross overall production of all crops is greater than 1750 KGs
Then
Probability of 'Non-poor' is 94% and 'Poor' is 6% (Accuracy = 0.94)

Class prediction is taken as **Non-poor**

Rule 2 (covers 293 cases)

If
total area of land owned is more than 1.62 acres and
gross overall production of all crops is greater than 1750 KGs
Then
Probability of 'Non-poor' is 94% and 'Poor' is 6% (Accuracy = 0.94)

Class prediction is taken as **Non-poor**

Rule 3 (covers 257cases)

If
at least one (or more) male member currently has a primary occupation and
total nos. of members is at most three
Then
Probability of 'Non-poor' is 72% and 'Poor' is 28% (Accuracy = 0.72)

Class prediction is taken as **Non-poor**

Rule 4 (covers 187 cases)

If
at most one male member is currently a student,
no male member is currently an agricultural labourer,
total area of operational land is more than 1.65 acres and
total area of irrigated land over which crops were grown in Season 1 is greater than 1.5 acres
Then
Probability of 'Non-poor' is 96% and 'Poor' is 4%(Accuracy = 0.96)

Class prediction is taken as **Non-poor**

Rule 5 (covers 88 cases)

If

no male member is aged between 0 and 1 years,
dependency ratio(dependants/working population) is at most 1.4,
no female member is currently an agricultural labourer,
total weight of all crops produced in Season 1 was more than 387.5 KGs and
total area of land over which plantains were grown is greater than 0.33 acres

Then

Probability of 'Non-poor' is 94% and 'Poor' is 6% (Accuracy = 0.94)

Class prediction is taken as **Non-poor**

Rule 6 (covers 75 cases)

If

total amount of chit fund is greater than Rs. 15000

Then

Probability of 'Non-poor' is 92% and 'Poor' is 8% (Accuracy = 0.92)

Class prediction is taken as **Non-poor**

Rule 7 (covers 72 cases)

If

no male member in the past was primarily occupied in trade and commerce[48]
at most one male member is currently a student,
no female member was a student in the past,
at most two female members currently have a primary occupation,[49]
the value of total Punjai Rain Fed (PRF) land owned is greater than Rs 32500,
value of total land leased out is at most Rs.25000 and
total amount of chit fund is at most Rs. 15000

Then

Probability of 'Non-poor' is 93% and 'Poor' is 7% (Accuracy = 0.93)

Class prediction is taken as **Non-poor**

Rule 8 (covers 79 cases)

If

at least one male member had a miscellaneous primary occupation in the past
and
at most two female members currently have a primary occupation

Then

Probability of 'Non-poor' is 73% and 'Poor' is 27% (Accuracy = 0.73)

Class prediction is taken as **Non-poor**

Rule 9 (covers 54 cases)

---

[48] Fishing has sometimes been included in this category and sometimes under miscellaneous

[49] This excludes female members who are primarily involved in housework, childcare and baby-sitting

If
no female member is aged between 0 to 4 years,
no male member is currently primarily occupied in the service sector,
no land is leased in,
own at least one oil engine which has some estimated monetary value and
the total number of members is at most five,
Then
Probability of 'Non-poor' is 84% and 'Poor' is 16% (Accuracy = 0.84)

Class prediction is taken as **Non-poor**

Rule 10 (cover 48)

If
at most one male member is educated up to the primary level[50]
no female member has in the past been primarily occupied in cultivation,[51]
no male member is ill or handicapped and
business assets in village have an estimated monetary value greater thanRs.1500
Then
Probability of 'Non-poor' is 92% and 'Poor' is 8% (Accuracy = 0.92)

Class prediction is taken as **Non-poor**

Rule 11 (cover 44)

If
at least one (or more) male member is aged between 35 and 49 years,
no male member is currently primarily occupied in the service sector,
at least one male member currently has a primary occupation,
total area of irrigated land over which crops were grown in Season 1 is at most 1.5 acres and
total number of members is less than or equal to 3
Then
Probability of 'Non-poor' is 87% and 'Poor' is 13% (Accuracy = 0.87)

Class prediction is taken as **Non-poor**

Rule 12 (cover 35)

[50] Primary level indicates up to class 7 usually aged around 12-13 years

[51] Cultivation includes own and tenancy cultivation

If
no male member is aged between 10 and 14 years,
no female member is aged lesser than 5 years
at least one male member is currently primarily occupied in the service sector,
at most two female members currently have a primary occupation
value of total land leased out is at most Rs.25000,
the estimated monetary value of plough bullocks is at most Rs 1,500 and
there are no business assets inside the village
Then
Probability of 'Non-poor' is 87% and 'Poor' is 13% (Accuracy = 0.87)

Class prediction is taken as **Non-poor**

Rule 13 (cover 32)

If
at most one male member is aged lesser than 10 years,
at least one male member is currently primarily occupied in construction work
and
total number of members is less than or equal to five
Then
Probability of 'Non-poor' is 85% and 'Poor' is 15% (Accuracy = 0.85)

Class prediction is taken as **Non-poor**

Rule 14 (cover 30)

If
at most one female member is working and
value of total land leased out is more than Rs.25000
Then
Probability of 'Non-poor' is 91% and 'Poor' is 9% (Accuracy = 0.91)

Class prediction is taken as **Non-poor**

Rule 15 (cover 27)

If
at most one female member is dependent
at least one male member had a miscellaneous primary occupation in the past,
no male member is currently an agricultural labourer,
at least one male member currently has a miscellaneous primary occupation,
at most one female member was an agricultural labourer in the past and
at most two female members currently have a primary occupation
Then
Probability of 'Non-poor' is 97% and 'Poor' is 3% (Accuracy = 0.97)

Class prediction is taken as **Non-poor**

Default rule

Default rule

If

The above conditions are not satisfied

Then

Class prediction is taken as **Poor**

6.3 Appendix III

Outlined below is the mechanism by which the alteration of the costs of errors affects rule formation. Details can be obtained from Quinlan, 1998.

When constructing the rule set, See5 obtains the Rule Set that has the lowest possible cost associated with its errors (rather than just lowest possible number of errors). When the two types of errors are considered equal (i.e. cost is equal), the performance of the Rule Set obtained is summarised by counting the errors (which in this case are the same as counting the cost) and obtaining the error rate. The 'cost' associated with a classification error can however be altered so that the cost of making one type of error can be considered 'x' times higher or lower than that of the other. This cost can be automatically included in the analysis when the Tree and Rule Set are being C. See5 then tries to obtain the Tree and Rule Set with the lowest possible cost. The manner in which this affects the Tree and Rule Set obtained is best illustrated by an example.

In our study when the cost of error Potentially incorrect non-Beneficiaries (IN i.e. 'Non-poor' household predicted as 'Poor') and Potentially incorrect beneficiaries (IB or 'Poor' household predicted as 'Non-poor') was considered equal the Rule Set obtained had the following 2×2 10 fold cross-validation table

Table A 'leave 10 out' cross validation 2×2 table on training set

		Predicted class		
		Poor	Non-poor	
Declared class	Poor	596	148	744
	Non-poor	213	543	756
		809	691	1500

Estimated predictive error of the rule set = (148 + 213)/1500 = 24%

The cost of the two errors IN and IB being equal implies that IN=IB=1. The total cost therefore = the total number of errors =213 + 418 = 361

Consider however e.g. the situation where the cost of making the error IN is considered 16 times that of making error IB. The Rule Set obtained has the following 10 fold cross validation 2×2 Table

Table B 'leave 10 out' cross validation 2×2 table on training set

		Predicted class		
		Poor	Non-poor	
Declared class	Poor	279	465	744
	Non-poor	73	683	756
		352	1148	1500

If we present the errors in Tables A and B here in terms of the costs, this can be expressed as follows:

Total Costs calculated for **Table A** when IN and IB have equal costs and when IN costs 16 times that of IB:

Cases	Predicted Class	Declared Class	Cost of errors (at cost IB=IN =1)	Cost of errors (at cost IN = 16 IB)
148	Non-poor	Poor	148	148
213	Poor	Non-poor	213	3408 (213×16)
			361	3556

Total Costs calculated for **Table B** when IN and IB have equal costs and when IN costs 16 times that of IB:

Cases	Predicted Class	Declared Class	Cost of errors (at cost IB=IN =1)	Cost of errors (at cost IN = 16 IB)
465	Non-poor	Poor	465	465
73	Poor	Non-poor	73	1168 (i.e. 73 × 16)
			438	1633

At equal costs, the Rule Set that corresponds to Table A had lower total cost and was thus selected. When the cost of IN was 16 times that of IB however, the Rule Set which corresponds to cross-validation Table B had lower total cost (though a higher number of total errors) and would presented as the output by See5.